Proceedings of Fonetik 2013

Edited by
Robert Eklund

Linköping University

Proceedings of Fonetik 2013

The XXVIth Annual Phonetics Meeting
12–13 June 2013, Linköping University
Linköping, Sweden

Studies in Language and Culture
no. 21

Robert Eklund, editor

Linköping University

Conference website: www.liu.se/ikk/fonetik2013
Proceedings also available at: http://roberteklund.info/conferences/fonetik2013

Cover design and photographs by Robert Eklund
Photo of Claes-Christian Elert taken by Eva Strangert on the occasion of his 80th birthday

Proceedings of Fonetik 2013, the XXVIth Swedish Phonetics Conference
held at Linköping University, 12–13 June 2013
Studies in Language and Culture, no. 21
Editor: Robert Eklund
Department of Culture and Communication
Linköping University
SE-581 83 Linköping, Sweden

ISBN 978-91-7519-582-7
eISBN 978-91-7519-579-7
ISSN 1403-2570

Printed by LiU-Tryck, Linköping, Sweden, 2013

This conference is dedicated to Professor Claes-Christian Elert
on the occasion of his 90th birthday, and for his outstanding
contributions to Swedish phonetics

Preface

This volume contains the contributions to Fonetik 2013, the XXVIth Swedish Phonetics Conference, held at Linköping University, Sweden, 12–13 June 2013, and was organized as a collaboration between the Department of Culture and Communication, Linköping University, and the Department of Clinical and Experimental Medicine, Division of Speech and Language Pathology, Linköping University.

The papers appear in alphabetical order, sorted by first author's last name. An author index is found on page 89.

Only a limited number of copies of this publication have been printed for distribution among the authors and those attending the conference. For access to electronic versions of the contributions and the entire proceedings, the conference homepage is found at:

http://www.liu.se/ikk/fonetik2013

The conference proceedings are also available at:

http://roberteklund.info/conferences/fonetik2013

The organizers would like to thank all contributors to the proceedings. We would also like to extend our thanks to *Fonetikstiftelsen* for generous financial support. Finally, we would thank the following people, whose help is very appreciated: Gunilla Christiansen, Carin Franzén, Nigel Musk, Anna F. Söderström and Linnea Björk Timm.

Linköping, May 2013

Robert Eklund, Agnese Grisle, Jan Anward, Charlotta Plejert, Christina Samuelsson and Simon Sundström

Previous Swedish Phonetics Conferences (from 1986)

I	1986	Uppsala University
II	1988	Lund University
III	1989	KTH Stockholm
IV	1990	Umeå University (Lövånger)
V	1991	Stockholm University
VI	1992	Chalmers and Göteborg University
VII	1993	Uppsala University
VIII	1994	Lund University (Höör)
—	1995	(XIIIth ICPhS in Stockholm)
IX	1996	KTH Stockholm (Nässlingen)
X	1997	Umeå University
XI	1998	Stockholm University
XII	1999	Göteborg University
XIII	2000	Skövde University
XIV	2001	Lund University
XV	2002	KTH Stockholm
XVI	2003	Umeå University (Lövånger)
XVII	2004	Stockholm University
XVIII	2005	Göteborg University
XIX	2006	Lund University
XX	2007	KTH Stockholm
XXI	2008	University of Gothenburg
XXII	2009	Stockholm University
XXIII	2010	Lund University
XXIV	2011	KTH Stockholm
XXV	2012	University of Gothenburg

Table of contents

Extracting and analysing co-speech head gestures from motion-capture data

Simon Alexanderson, David House & Jonas Beskow
Department of Speech, Music and Hearing, KTH, Stockholm, Sweden

Abstract

This paper reports on a method developed for extracting and analyzing head gestures taken from motion capture data of spontaneous dialogue in Swedish. The head gestures were extracted automatically and then manually classified using a 3D player which displays time-synced audio and 3D point data of the motion capture markers together with animated characters. Prosodic features were extracted from syllables co-occurring with a subset of the classified gestures. The head gestures show considerable variation in temporal synchronization with the syllables, while the syllables generally show greater intensity, higher F0, and greater F0 range when compared to the mean across the entire dialogue.

Introduction

Along with intonation there are a wealth of visual gestures, including head, facial and body movements, which co-occur with speech adding emphasis and prominence to portions of the utterance and contributing to the flow of the dialogue. Beat gestures such as rapid hand movement as described by McNeill (2005) are particularly interesting in this regard as they coincide and appear to be synchronized with prosodic and intonational peaks related to prominence. They also share the same prominence-lending function as the focal or sentence accent, but they can be repetitive marking the stress or rhythmical structure of an utterance.

In terms of timing and synchronization, there are many similarities between intonational gestures and visual gestures produced in accompaniment with speech. Both intonation and visual gestures are free to vary across the vowels and consonants of the segments. In intonation, however, this variation is restricted by the specific patterns used by a language to signal meaning in spoken interaction.

If we wish to study the timing of gestures in the same way as we approach timing in intonation, we currently lack an established methodology to extract and analyze gestures, especially gestures occurring in spontaneous dialogue. The Spontal corpus of Swedish dialogue provides a rich database as a point of departure for testing gesture extraction and analysis methodology. The database, containing more than 60 hours of unrestricted conversation in over 120 dialogues between pairs of speakers is comprised of high-quality audio and video recordings (high definition) and motion capture for body and head movements for all recordings (Edlund et al., 2010).

The progression and timing of the motion of a head nod can be described in much the same terms as an intonational excursion. However, many speakers move their heads extensively while speaking, and manual annotation of head-gestures in spontaneous dialogue involves a number of difficulties. Among the sources of disagreement are segment boundaries and location of maximum extent. Gestures may be multifunctional and involve simultaneous rotations around several axes. In this study we present a semi-automatic approach to head-gesture annotation, in which the main goal is to test its viability and potential for annotation of gesture data on a large scale, such as is represented by the Spontal database.

Method

To overcome some of the difficulties of head-gesture annotation we are developing and testing a semi-automatic annotation procedure consisting of two steps. First an automatic head-gesture segmentation algorithm is applied to the motion capture data and then the segments are manually classified by the annotators.

In addition to gesture annotation we also processed the audio files of the speakers and generated pitch and intensity data, talk vs. no-talk segmentation and syllable segmentation. This was done to be able to investigate the relationship between the head-gestures and prosodic features.

In this exploratory stage we chose one of the dialogues from the Spontal database where the speakers demonstrated a relatively large number and variety of head-movements. The participants were a male and female who did not know each other.

Automatic segmentation of head nods

A simplistic segmentation approach was used for head-nod segmentation. The head orientations were calculated from three markers attached to the headbands of the speakers and expressed in an Euler angle form. We then calculated the angular velocity of the pitch component as the basis for the segmentation. During a head nod the angular velocity follows an oscillatory movement during a limited time period, and we use its local extreme values as segment boundaries. A segment is defined as a maximum velocity followed by a minimum or vice versa. Two thresholds are used in this process. The first enforces the peak velocity to be over a specified value, thus prohibiting small movements caused by noise to be interpreted as nods, and the second enforces the nod segments to be shorter than a specified duration. Using the velocity peaks as segment boundaries has some desirable features. During head-nods there are rapid changes in angular velocity causing clear detectable spikes in the data. It also naturally splits repeated head-nods into a consecutive sequence of down-up (nod) and up-down (jerk) segments, which fits well with the MUMIN multimodal annotation scheme proposed by Allwood et al. (2007).

For our data, the minimum peak velocity threshold was empirically set to a value of 0.0015 radians/s, which was the lowest value before noise in the data would be manifested as segments. The maximum segment duration threshold was set to 1000 ms.

In the current study we were interested in gestures synchronized with speech and especially gestures with beat-function produced in companion with stressed syllables. In order to narrow our search space we discarded all segments occurring while the subject was not speaking and further all nods in the up-down order, leaving all down-up nods occurring during speech as our candidate gestures. The segmentation of talk vs. no-talk was performed with an automatic speech activity detection algorithm (Heldner et al., 2011).

Manual classification

After running the automatic processing, the resulting segments were examined and manually classified by two annotators. To make the classification, the annotators viewed each segment in a specially designed 3D player which plays time-synchronized audio and displays 3D point data of the motion capture markers together with animated characters following the 3D marker motion. As expected, the segments from the automatic process did not only contain unambiguous beat gestures, but also gestures with other functions co-occurring with speech (McClave, 2000). Such other functions were feedback, confirmation, word or phrase intensification and listing of lexical items. Moreover, some of the extracted gestures did not appear to co-occur with a stressed syllable.

Therefore, an annotation scheme was devised with two main queries: Q1, "Is there a clear nod in synchrony with a stressed syllable?" and Q2, "Is the nod multifunctional?" If the answer to the first query is positive the second query is also answered. This scheme resulted in three categories: 1. No clear nod in synchrony (no sync), 2. A clear nod with a beat function (beat-function), and 3. A clear nod which is multifunctional (multi-function).

Prosodic features

The pitch and intensity curves were extracted from the audio signals of the speakers using the SNACK toolkit (Sjölander & Beskow, 2000). Also syllable boundaries and nuclei were derived by applying Mermelstien's convex-hull algorithm (Mermelstien, 1975).

After gesture- and prosodic feature extraction was performed, we determined which syllable was closest in time to the maximum rotation of the nod. The time difference between each gesture and the start and nucleus of its closest syllable was then calculated. Also pitch and intensity properties of the closest syllable were compared with mean values across the total dialogue.

Results

The automatic segmentation algorithm was applied on the 20 minute dialog, extracting 64 nod-segments for speaker 1 (male) and 150 segments from speaker 2 (female). The manual classification by the two annotators resulted in a 69% and 65% agreement for the head-nods of speaker 1 and speaker 2 respectively. As is displayed in *Table 1* the annotators showed greatest agreement for the category with no syllable-synchronization. Less agreement was obtained from the categories beat-function and multi-function. Annotator 1 perceived more of the nods having beat-function while annotator 2 perceived more having multi-function.

Table 1: Results of the manual annotation showing class and agreement. The first number in each pair is the male speaker, the second is the female.

Class	Annotator 1	Annotator 2	Agreement
No sync	23/44	15/41	13/29
Beat-function	9/74	11/66	4/50
Multi-function	32/32	38/43	27/18
Total	64/150	64/150	44/97

Figure 1 shows the durations of the segments in the different categories for speaker 1 and speaker 2 for those gestures for which the annotators agreed.

Note that the segment length is the part of the nod between the peak velocities of the downwards and upwards phase as described earlier. The results show a tendency for the multi-functional nods to be shorter than those with a beat function. The nods classified as non-synchronous showed greater temporal variation than the other categories.

The subset of gestures annotated as having a beat function for the female speaker was analyzed in terms of timing related to its closest syllable. Only the gestures from the female speaker were analyzed due to the small number of beat gestures annotated for the male speaker. *Figure 2* shows the time difference between two different anchor-points of the syllable (onset and nucleus) and three different phases of the nod: peak velocity of the downward phase (p1), max rotation (p2) and peak velocity of the upward phase (p3). The timing relationship between the gesture and the syllable does not seem to be influenced by the choice of syllable anchor-point. The timing relationships show a considerable amount of variation regarding the question of gesture synchronization with the syllable.

When compared with mean values across the total dialogue, the syllables closest to the annotated beat nods generally showed greater integrated intensity, higher F0 at the nucleus, and greater F0 range as shown in *Table 2*.

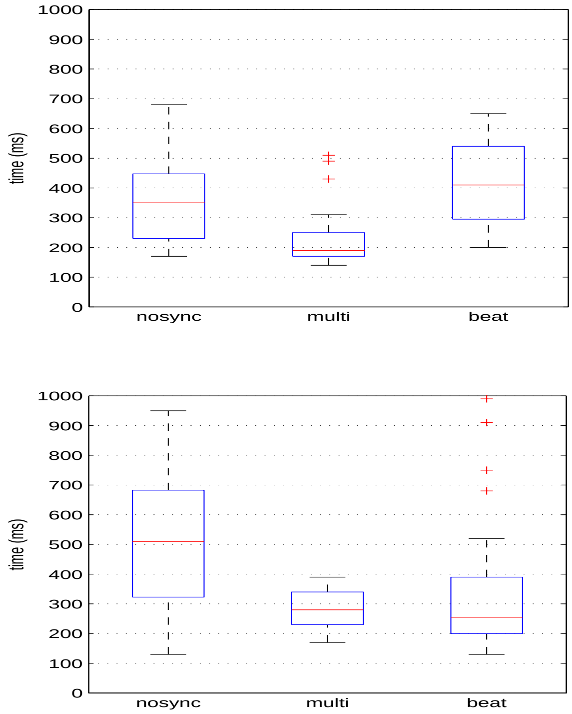

Figure 1: Durations of the agreed nods in the classes for speaker 1 (top) and speaker 2 (bottom).

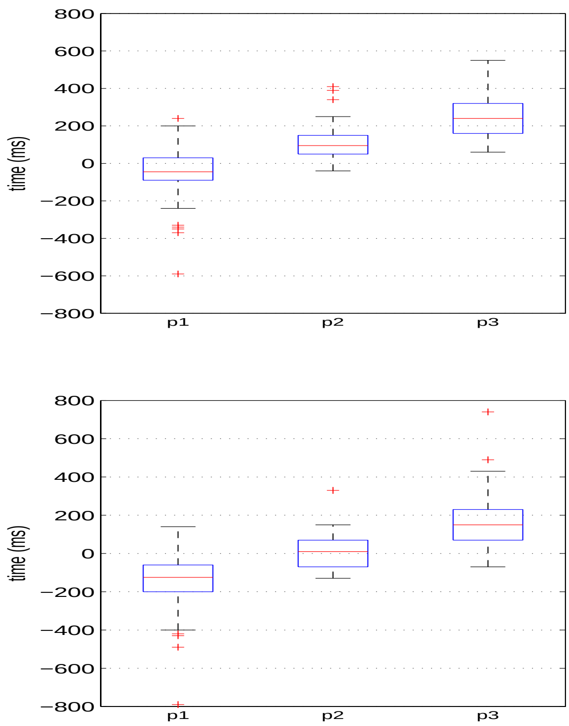

Figure 2. Timing of syllable anchor-points, onset (top) and nucleus (bottom), with respect to three different phases of the nod: peak velocity of the downward phase (p1), max rotation (p2) and peak velocity of the upward phase (p3).

Table 2. Comparison of syllable features

	integrated intensity (dBs)	F0 at nucleus (Hz)	F0 range (Hz)
Mean closest syllable	13.2	217	77
Mean across the total dialogue	11.1	180	67

Discussion

The goal of this study was primarily to develop and test a new method for extracting and annotating gesture data. While head gestures, and in particular relatively small gestures, have been problematic for manual annotation schemes, the semi-automatic method tested here shows promise for the selection and fast annotation of multimodal data.

This process is a starting point for further work in the field of automatic recognition and classification of multimodal communication. Given the fact that non-verbal and verbal communication are tightly coupled, the motion data may provide important and robust features for machine-learning techniques. In this study we started an investigation along this path by analyzing features for prominence detection and their coupling to beat gestures. This method may also prove useful in analyzing features related to other communicative functions such as feedback and turn-taking.

While the results concerning the analysis of the characteristics of the head nods and their timing must be seen as quite preliminary due to the small sample and very limited classification categories, the timing results are consistent with those results reported in Leonard and Cummins (2011) for hand and arm beat gestures. More available data and the development of automatic methods and tools should better enable us to compare and evaluate results such as these.

Acknowledgements

The work reported here is carried out within the projects: “Timing of intonation and gestures in spoken communication,” (P12-0634:1) funded by the Bank of Sweden Tercentenary Foundation, and “Large-scale massively multimodal modelling of non-verbal behaviour in spontaneous dialogue,” (VR 2010-4646) funded by the Swedish Research Council. A longer version of this paper will be presented at the Tilburg Gesture Research Meeting in June 2013.

References

Allwood, J., L. Cerrato, K. Jokinen, C. Navarretta & P. Paggio. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation* 41:273−287.

Edlund, J., J. Beskow, K. Elenius, K. Hellmer, S. Strömbergsson & D. House. 2010. Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture. In: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner & D. Tapias (eds.), *Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10).* Valetta, Malta, 2992−2995.

Heldner, M., J. Edlund, A. Hjalmarsson & K. Laskowski. 2011. Very short utterances and timing in turn-taking. In: *Proceedings of Interspeech 2011*. Florence, Italy, 2837−2840.

Leonard, T. & F. Cummins. 2011. The temporal relation between beat gestures and speech. *Language and Cognitive Processes* 26: 1457−1471.

McClave, E. 2000. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics* 32:855−878.

McNeil, D. 2005. *Gesture and thought.* Chicago: The University of Chicago Press.

Mermelstien, P. 1975. Automatic segmentation of speech into syllabic units. *Journal of the Acoustical Society of America* 58:880−883.

Sjölander, K. & Beskow, J. 2000. WaveSurfer – an open source speech tool. In: B. Yuan, T. Huang & X. Tang (eds.), *Proceedings of ICSLP 2000, 6th International Conference on Spoken Language Processing*. Beijing, China, 464−467.